REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12-08-2012	Final Technical Report	03-01-2009 - 29-01-2012

4. TITLE AND SUBTITLE
Geometric Networks Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0143

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Gunnar Carlsson, Mike Mahoney.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Stanford University
320 Panama Street, Bambi Modular
Stanford, CA, 94305-4100

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N. Randolph, Ste.325
Arlington Virginia, 22203
Dr. Robert Bonneau/RSL

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0973

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project has made numerous contributions to our understanding of large and complex networks. One is the construction of bounds on the complexity of classification problems carried out on graphs, rather than in Euclidean space. In this case, the machne learning is performed in very high variance environments, so traditional methods based on bounding the Vapnik–Chervonenkis dimension fail, and a theorem bounding the annealing entropy for classifiers in a Hilbert space is proved as a replacement. Another contribution is a locally biased analogue of the second eignevector of the Laplacian metrix attached to a graph. It is demonstrated that this eigenvector can highlight local properties of data graphs effectively. Also, Gromov hyperbolicity of networks constructed using various generative schemes were studied, and it was observed that these schemes did not substantially improve the hyperbolicity of the networks in question. Further, a new family of graphs were introduced, the "ringed tree", was introduced, as a generalization of a tree. It provides an analogue for graphs of the hyperbolic plane. Another contribution as the analysis of what obj

15. SUBJECT TERMS
Complex networks, machine learning, spectral clustering.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gunnar Carlsson
U	U	U	U	5	19b. TELEPHONE NUMBER *(Include area code)* 650 723 2224

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Final Report on AFOSR FA9550-09-1-0143: Geometric Networks Analysis

August 12, 2012

1 Introduction

The main goal of the project was to develop and use algorithmic tools with a geometric underpinning to analyze large networks. For example, the Internet, social networks, connectivity properties between genes or proteins in a biological cell, and other types of data that may be viewed as finite metric spaces can all be fruitfully modeled as a network or a graph $G = (V, E)$ consisting of nodes and edges. A general property of many of these networks is that there is some sort of very local structure (*e.g.*, clustering coefficients in so-called "small world" models, or degree distribution in so-called "power law" models) that is meaningfully geometric (in the sense of having a low-dimensional Euclidean geometry or other "nice" geometry associated with it), and the graph is relatively-unstructured otherwise. When working with networks with millions (or more) of nodes, algorithmic considerations are central; but those networks are typically extremely sparse and noisy, and so understanding the statistical properties underlying those algorithmic tools is also central. Developing and using principled algorithmic tools to extract and exploit this structure, as well as using the statistical and geometric properties underlying those tools to certify in a reliable manner when such structure is not present, were primary goals of the research.

Output of the research includes numerous conference and journal publications, a series of tutorials and keynotes that were the basis for two overview articles as well as a soon-to-be-completed monograph, and several other projects that will soon be completed. In more detail, the published results of the supported research include:

- [3] is a much longer journal version of the initial work that motivated the supported research. In addition, there was another conference paper that described in more detail empirical properties of approximation algorithms for network community detection [4]. Among other things, this latter work included the first empirical demonstration of statistical regularization implicit in scalable worst-case approximation algorithms for the intractable graph partitioning problem for large real-world informatics graphs.

- [9] proved bounds on the sample complexity of maximum margin classifiers when the magnitude of the entries in the feature vector decays according to a power law and also when learning is performed with the so-called Diffusion Maps kernel on general, *i.e.*, non-manifold-like, graphs. The key property of both of these situations is that the learning is done in potentially very high-variance environments, and so existing methods based on bounding the VC dimension fail to give nontrivial results. Thus, the results required a more general theorem on bounding the annealed entropy of gap-tolerant classifiers in a Hilbert space (which generalizes to the case when the margin is measured with respect to more general Banach space norms).

- [11] introduced a locally-biased analogue of the second eigenvector of the Laplacian matrix, and it demonstrated the usefulness of that vector at highlighting local properties of data graphs in a semi-supervised manner. In particular, it provides an optimization formulation of locally-biased spectral methods that can be used to find clusters in very large networks.

- [10] addressed the question of what is the regularized optimization objective that an approximation algorithm is exactly optimizing; and it provided a precise answer in the context of three random-walk-based procedures (namely, those used by local spectral methods) for computing approximations to the second eigenvector of a graph Laplacian. [12] then followed up this work to provide a statistical interpretation of these random-walk-based approximation algorithms. The interpretation provided a precise statistical and geometric sense in which random-walk-based approximation algorithms can be interpreted as regularized estimates of the pseudoinverse of the graph Laplacian.

- [7] and then [8] addressed algorithmic and statistical perspectives on large-scale data analysis. They were articles written to accompany two invited keynote lectures, that were subsequently given in many additional venues, and they described several case-studies of how to use exploit the geometric and statistical properties underlying worst-case approximation algorithms in practical data applications.

- [2] empirically evaluated and described the phenomenon of localization on low-order eigenvectors, *i.e.*, those eigenvalues not associated with extremal eigenvalues, of data matrices, *i.e.*, adjacency matrices and Laplacians of networks. In many cases, this phenomenon, which is typically simply assumed not to exist in most machine learning theory, can be interpreted in terms of tensor-product-like behavior that is a consequence of the time evolution of the underlying network.

- [1] analyzed the hyperbolicity of small-world networks and tree-like graphs using Gromov's notion of δ-hyperbolicity. Among other things, it showed that popular small-world constructions do not substantially improve the δ-hyperbolicity of the network, even when the rewiring precess is chosen such that the network is navigable in a decentralized manner; and it introduced the idea of a ringed-tree as a softened version of a binary tree that is quasi-isometric to the Poincare disk.

In addition to these already-published results, the supported research included the presentation of a series of tutorials on "Geometric Tools for Graph Mining of Large Social and Information Networks." These were originally given at ICML 2010 and KDD 2010, and they have been reinvited and given at more than a dozen venues since then. In addition, these presentations are the basis for a monograph on "Geometric structure in large informatics graphs" that is scheduled to be completed by the end of the current academic year [5] and that will be published NOW Publishers' Foundations and Trends in Machine Learning series (where Mahoney's other recently-completed monograph on "Randomized algorithms for matrices and data" was just published [6].

The supported research has also included several projects with a Stanford graduate student (supported on another grant) and collaborators at Oak Ridge National Labs and elsewhere. These projects include: empirically evaluating tree-like structure in real networks by using the complementary approaches of δ-hyperbolicity (which characterizes tree-like-ness in terms of metric structure) and tree decompositions (which characterize tree-like-ness in terms of cut structure) in order to identify and exploit in a scalable and robust way meaningful tree-like structure in large social and information networks; providing theoretical characterizations of relationships between δ-hyperbolicity and notions such as tree-width and tree-length; using k-core-based algorithms to identify local structure in large social and information networks in a manner that is complementary to the use of local spectral methods; and extending Poincare disk-based hyperbolic visualization methods that have been developed recently for Internet routing applications to visualizing geometric structure at several size scales in large social and information networks. Although this supported research has not yet been completed, it will be written up and published in appropriate venues when it is completed.

References

[1] W. Chen, W. Fang, G. Hu, and M. W. Mahoney. On the hyperbolicity of small-world networks and tree-like graphs. Technical report. Preprint: arXiv:1201.1717 (2012).

[2] M. Cucuringu and M. W. Mahoney. Localization on low-order eigenvectors of data matrices. Technical report. Preprint: arXiv:1109.1355 (2011).

[3] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009. Also available at: arXiv:0810.1355.

[4] J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, 2010.

[5] M. W. Mahoney. *Geometric structure in large informatics graphs.* Manuscript in preparation.

[6] M. W. Mahoney. *Randomized algorithms for matrices and data.* Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.

[7] M. W. Mahoney. Algorithmic and statistical perspectives on large-scale data analysis. In U. Naumann and O. Schenk, editors, *Combinatorial Scientific Computing*, Chapman & Hall/CRC Computational Science, pages 000–000. CRC Press, 2012.

[8] M. W. Mahoney. More algorithmic and statistical perspectives on large-scale data analysis. In *Proceedings of the 31th ACM Symposium on Principles of Database Systems*, pages 000–000, 2012.

[9] M. W. Mahoney and H. Narayanan. Learning with spectral kernels and heavy-tailed data. Technical report. Preprint: arXiv:0906.4539 (2009).

[10] M. W. Mahoney and L. Orecchia. Implementing regularization implicitly via approximate eigenvector computation. In *Proceedings of the 28th International Conference on Machine Learning*, pages 121–128, 2011.

[11] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi. A spectral algorithm for improving graph partitions with applications to exploring data graphs locally. Technical report. Preprint: arXiv:0912.0681 (2009).

[12] P. O. Perry and M. W. Mahoney. Regularized Laplacian estimation and fast eigenvector approximation. In *Annual Advances in Neural Information Processing Systems 25: Proceedings of the 2011 Conference*, 2011.